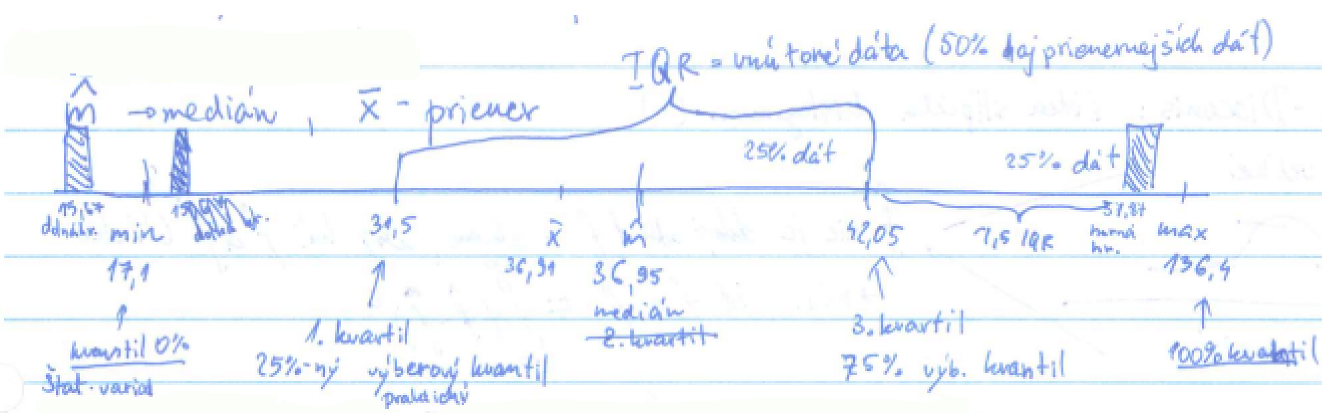
Počítačová štatistika

1976 – Bell labs S - statistics (narážka na C) 1988 – S+ – konečná implementácia S 1997 – R – Eobert Gentleman, Ross Ihaka, odlíšiť od S

R-ko

- attach: kópie stĺpcov v glob. namespace
- logické: & , |, ==, ! =, !
- order: vektor pozícií v poradí
- rank: pozície v usp. poli
- data.frame: matica objektov
- Ctrl+R – run
- vektory = stĺpce
- násobenie matíc: %*%
- $solve(A, \vec{v})$ → riešenie $A\vec{x} = \vec{v}$
- objekt$param – stĺpec
- NAN - not a number
- NA - not available = chýbali dáta
- imputačná technika – okus nahradiť NA nejakými dátami
- FACTORS – premenné, ktoré nie sú číselné, LEVELS OF FACTOR – hodnoty tejto premennej(set: M,F)
- table – kontingenčná tabuľka

Štatistiky a variability

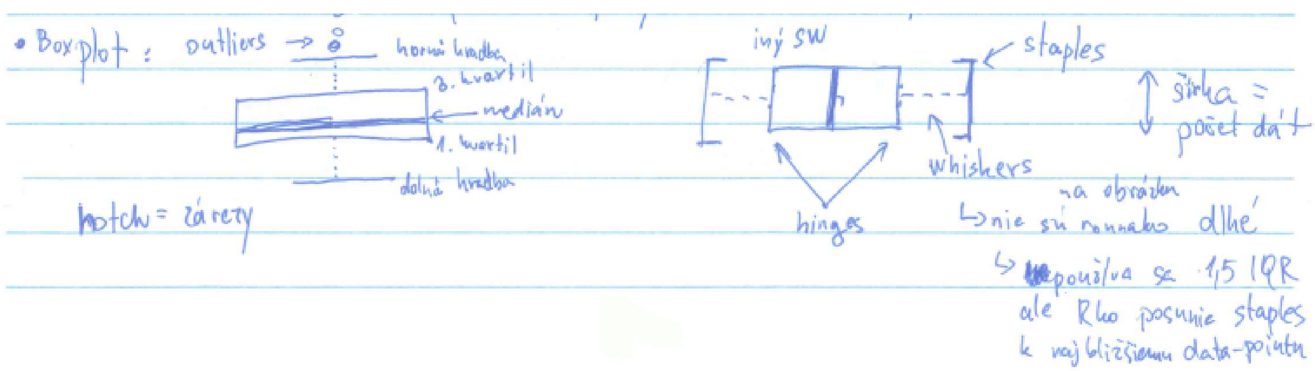


- S^2 – variabilita dát
  - význam má hlavne pri porovnávaní 2 vzoriek dát
  - v štat. na normalizácie
- IQR – interquartile range – uchopiteľnejšie ako S^2
- dolná hradba = 1. kvantil – 1.5*IQR : prečo 1.5? – určuje šírku hradieb – pre bežné dáta funguje[1]
- outliers → mimo hradieb → treba sa pozrieť na to, ako vznikol? → zaradiť & vyhodiť (chyba meranie)
  - ak ich necháme, robí sa analýza s out. aj bez out. samostatne

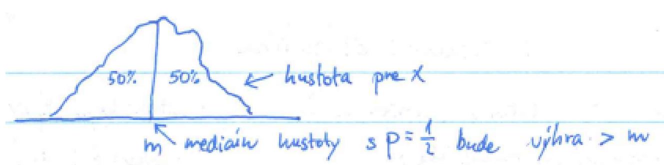
Obrázky

Lotéria: všetci si stavili číslo 0-999, tí čo si tipli vyžrebované si rozdelili peniaze

- Boxplot: outliers



Zárezy



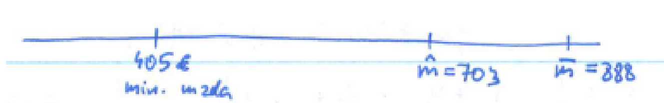
X – náhodná premenná

odhady pre m:

1. bodový odhad: $\hat{m}$ . . . medián z dát (výberový)
2. intervalový odhad: IS pre m: $\hat{m} \pm$ čosi

Boxploty sa vždy maľujú viaceré
Medián je robustný = odolný voči outlierom 5,1,1,2,1 → $\Phi = 2 \wedge \hat{m} = 1$

medián + priemer: Viac ako polovica Slovákov má podpriemernú mzdu:



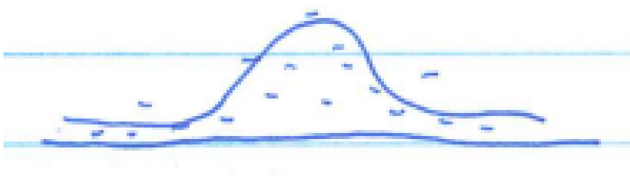
[1]nie sú príliš úzke ani široké

Norm. rozdelenie $N(0,1)$ $f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$
Cauchyho rozdelenie $f(x) = \frac{1}{2\pi} \frac{1}{1+x^2}$

Freedman-Diaconis: šírka stĺpčeka histogramu (n)
pre *a)* aj *b)*: $\hat{f}$ nie je dobrý odhad f – chceme, aby boli f a $\hat{f}$ blízko – min. vzd. f a $\hat{f} = \int_{-\infty}^{\infty} (\hat{f} - f)^2$

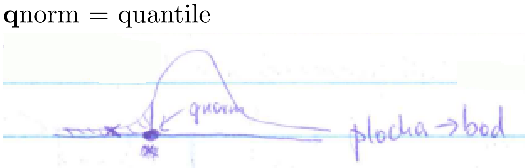
a) h je veľké

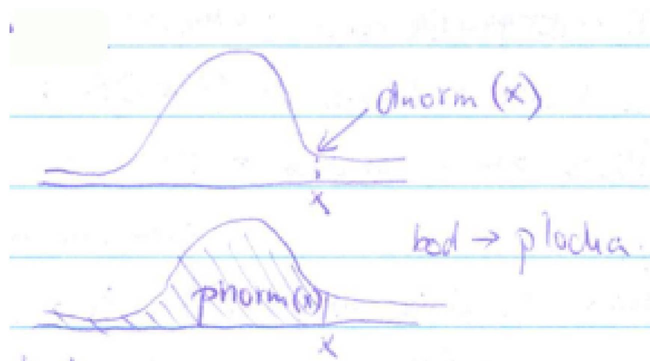
b) h je malé $\rightarrow$ pre veľkú triedu funkcií je riešením práve $h = 2 \cdot IQR \cdot n^{-\frac{1}{3}}$



Rozdelenia a hustoty v R

- `dnorm(x, mean=4, sd=30)`
  - **d**norm = density
  - d**norm** = norm. rozdelenie
  - mean=4, sd=3 ... hustota $N(4,9)$, dke 9 je σ^2
- `pnorm(7)` = distrib. f. v 7 = $F(7) = P[x < 7]$
  - **p**norm = probabilty
- `qnorm(0.09)` = 9% kvantil
  - **q**norm = quantile

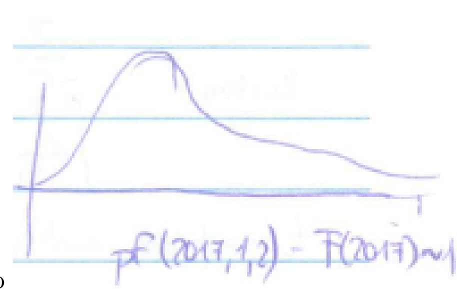




- `rnorm(103)` – vygeneruje sa $x_1, \ldots, x_{103} \sim N(0,1)$ a sú IID

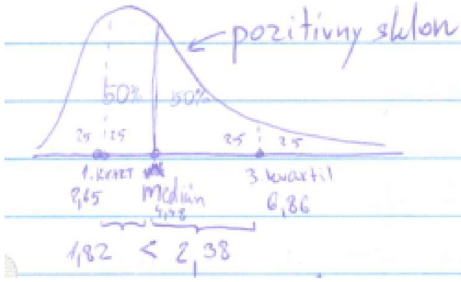
POZN: množstvo štat. metód vyžaduje ďalšie generovanie dát

- `norm` – normálne rozdelenie
- `t(..., df)` - Studentovo rozdelenie t_{df}, df = degrees of freedom
- `chisq(..., df=4)` – χ^2_4 – chi kvadrát
- `f(..., df1=2, df2=8)` – Fischerovo-Schnedekerovo



Sklon dát

Väčšina dát je z norm. rozdelenia, ale treba overiť!



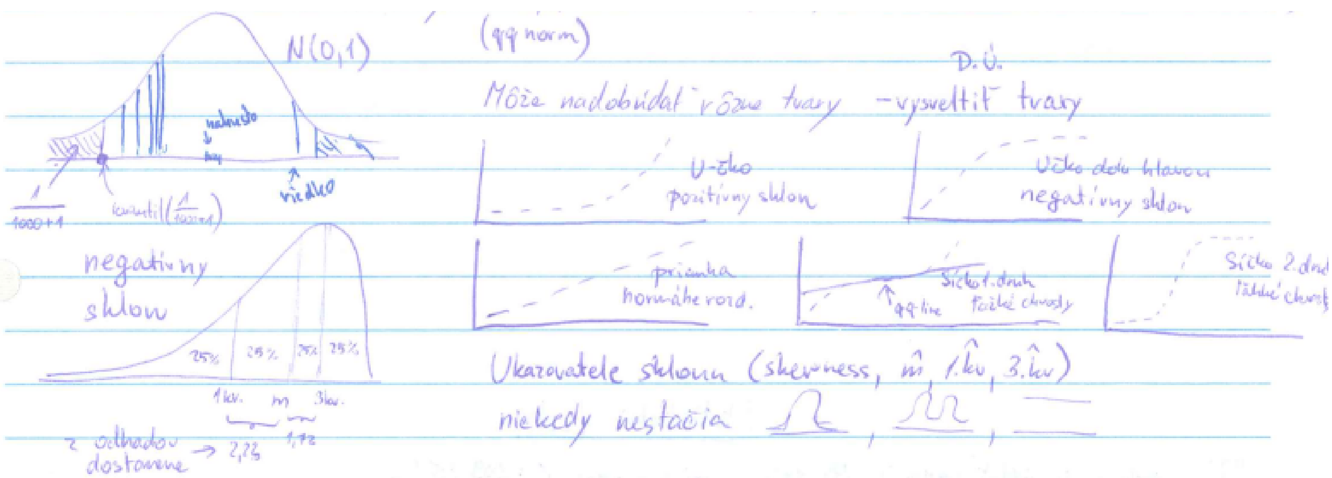
pozitívny sklon: $(3\hat{k} - \hat{m}) > (\hat{m} - 1\hat{k})$
(jednoduchšie, ale dáta treba usporiadať)

pozitívny sklon – ako to matematicky zistiť?

$$\text{skewness} := \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^{\frac{3}{2}}}$$

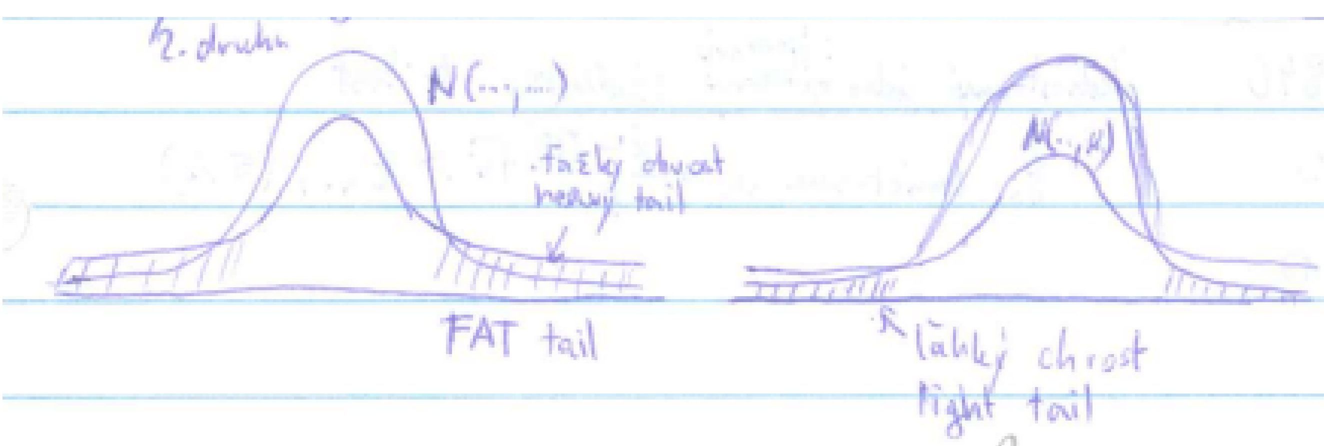
- $skewness > 0$ – pozitívny sklon
- $skewness = 0$ – symetrická hustota
- $skewness < 0$ – negatívny sklon

Q-Q plot (Quantile-Quantile plot – kvantilový diagram)
y-ové súradnice guľôčok sú usporiadané dáta $x_{(1)} < x_{(2)} < \ldots < x_{(1000)}$
x-ové súradnice kvantily $N(0,1)$ $\quad kvantil(\frac{1}{1000+1}) < kvantil(\frac{2}{1000+1}) < \ldots < kvantil(\frac{1000}{1000+1})$



qqline – priamka cez body $\frac{1}{4}$, $\frac{3}{4}$

- Esíčkový kv. diagram **1. druhu**: Dáta pochádzajú z hustoty, ktorá má ťažšie chvosty než $N(\ldots,\ldots)$, a teda špicatejší kopček než $N(\ldots,\ldots)$
- Esíčkový kv. diagram **2. druhu**: Dáta sú z rozdelenia s hustotou s ľahšími chvostami a mohutnejším kopčekom



KURTOSIS (vydutosť) – výberový keoficient špicatosti

$$\text{kurtosis} := \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^2}$$

- *kurtosis* > 3 – špicatejší kopček, ťažšie chvosty
- *kurtosis* $= 3$ – dáta sú z norm. rozdelenia
- *kurtosis* < 3 – mohutnejší kopček, ľahšie chvosty

teoretická kurtosis – získame ju z hustoty:

$$\frac{E[(X - E(x))^4]}{E[(X - E(x))^2]^2} = 3$$

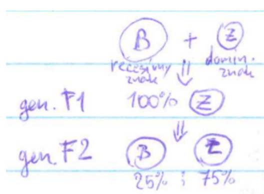
Rovná sa 3, bez ohľadu na μ a σ^2. Menovateľ je σ^{2^2}. Ešte treba ukázať, že čitateľ je $3\sigma^4$.
reálne dáta – hľadáme <u>hsutotu</u>, ktorá sa najviac podobá na naše dáta

- metóda: histogram (schodíkový odhad)
- <u>jadrové odhady</u> (kernel estimates) – v Rku density

Obrázok – histogram, gauss, kernel estimate $\rightarrow$ ak sa podobá kernel est na gaussa, pravdepodobne to bude norm. rozdelenie

Historické príklady falšovania dát

- franc. profesor – hoďťe mincou 1000 krát a odovzdajte zápis
  - pomer 500:500 sa im podarilo napodobniť
  - tí čo podvádzali nemali dlhé rovnaké sekvencie
- Mendel (brnenský mních, objavil zákony dedičnosti) – kríženie hrachov s bielymi a zelenými kvetmi
  - pozrel sa na dáta Fischer (štatistik) – zákon síce platil, ale experimenty (veľa hrahov) sú falšované
  - 100 hrachov na 1 políčku: 26:47, 25:75, 24:76 – skutočné pomery by boli ďalej od 25:75 – ukázal pomocou <u>testu dobrej zhody</u>
  - pravdepodobne za to môže pomocník: zatajil výsledky 85:15 $\rightarrow$ <u>CONFIRMATION BIAS</u> – zbieranie len dát, ktoré potvrdzujú hypotézu



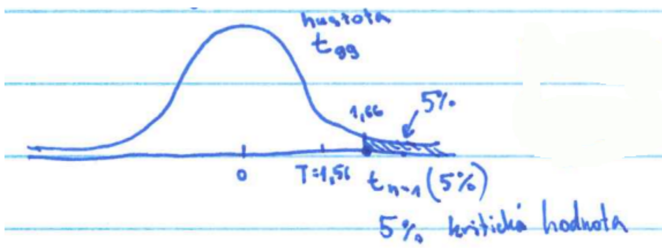
Štatistické testy

1878: verilo sa, že rýchlsoť svetla je 299 840, Michelsonov $\bar{X} = 299\,852,4$
$X_1, \ldots, X_{n=100} \sim N(\mu, \sigma^2)$, μ nepoznáme = skutočná rýchlosť svetla
Spravíme Studentov t-test

$$H_0 : \mu \leq 840 \qquad vs. \qquad H_1 : \mu > 840 \text{ (research hypothesis)}$$

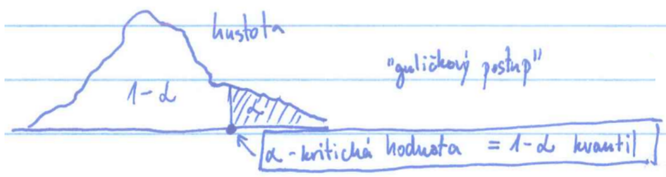
One-sided t-test

MZR: H_0 zamietame ak $\bar{X} >> 840$. Jednostranný súborový studentov t-test.



H_0 zamietame ak $T = \frac{\bar{X}-840}{S}\sqrt{n} > t_{n-1}(5\%)$

t-test v Rku: `t.test(Y, alternative=greater, mu=840)`
$t = T = \frac{\bar{X}-840}{S}\sqrt{n} = 1.5694$ chýba tu kritická hodnota $t_{99}(5\%)$ → Rko nevie robiť krit. hodnoty ako kvantily



$T \not> t_{99}(5\%)$ $\qquad t_{99}(5\%) = 1.66$
– hypotézu H_0 nezamietame

– Mechelspon nemôže tvrdiť, že $c > 299\ 840$, Studentov t-test zohľadnil

1. rozdiel 840 a 852 je dosť malý
2. vzorka 100 dát je malá
3. dáta sú rozdistribuované

p-value $\in < 0, 1 >$, $\qquad p = 5.98\%$ – plocha od $T \to \infty$
Ak p-value $< 5\%$, H_0 zamietame, ak p-value $> 5\%$, H_0 nezamietame

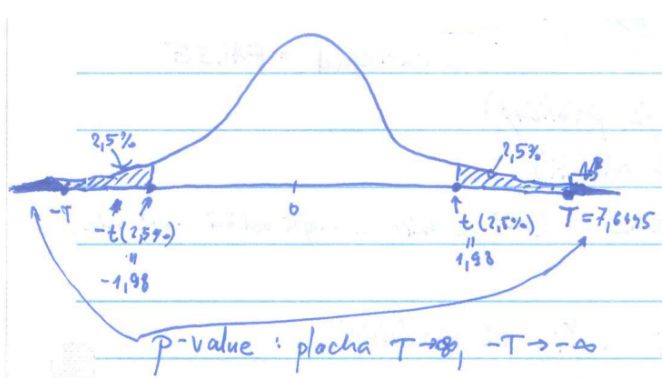
- "p-value" – porovnaj t a T
- "guľičkový postup" – porovnaj plochy 5% a $\int_T^\infty p(x)dx$

Two-sided test

TRUE: 299 792 km/h – Porovnáme Michelsonove dáta s realitou

$$H_0 : \mu = 792 \qquad vs. \qquad H_1 : \mu \neq 792$$

MZR: H_0 zamietame ak $\bar{X} >> 792$ alebo ak $\bar{X} << 792$.



$$\bar{X} >> 792 \text{ alebo } \bar{X} << 792$$
$$\bar{X} - 792 >> 0 \text{ alebo } \bar{X} - 792 << 0$$
$$T = \frac{\bar{X} - 792}{S}\sqrt{n} > t(2.5\%) \text{ alebo } < -t(2.5\%)$$
$$7.6445 > 1.98 \text{ alebo } 7.6445 < -1.98$$

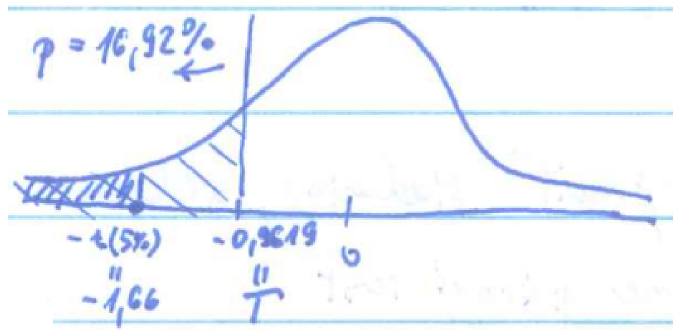
Zamietame H_0, problém! – systematická chyba merania, michelson dostáva dáta s $\mu > c$.

Ešte jeden 1-stranný t-test

V 1878 sa verilo, že NEWCOMBE: 860 (MICHELSON: 852.4 → bol bližšie k TRUE: 792)

$$H_0 : \mu \geq 860 \qquad vs. \qquad H_1 : \mu < 860$$

H_0 zamietame ak $\bar{X} << 860$



$$T = \frac{\bar{X} - 860}{S}\sqrt{n} \qquad < -t(5\%) = t(95\%)$$

$$-0.9619 \qquad > -1.66$$

H_0 zamietame, test nie je štatisticky významný

p-value $= p(\tau < -0.96) = F_T(-0.96)$

Dva dni meraní pre rýchlosť svetla

$X_1, \ldots, X_n \sim N(\mu_x, \sigma^2)$
$Y_1, \ldots, Y_n \sim N(\mu_y, \sigma^2)$
predpokladáme, že sigmy sú rovnaké, ale treba to overiť
μ_x a mu_y – skutočné rýchlsoti svetla v určitých dňoch:

$$H_0 : \mu_x = \mu_y \qquad vs. \qquad H_1 : \mu_x \neq \mu_y$$

Predtest: $H_0 : \sigma_x^2 = \sigma_y^2 \qquad vs. \qquad H_1 : \sigma_x^2 \neq \sigma_y^2$
MZR: H_0 zamietame ak $S_x^2 >> S_y^2$ alebo $S_x^2 << S_y^2$.
F-test: $\frac{S_x^2}{S_y^2} >> 1$ alebo $\frac{S_x^2}{S_y^2} << 1$ (var. test)
2. a 4. deň $\frac{S_x^2}{S_y^2} = F = 1.0377 \qquad p - value = 93\% \Rightarrow H_0$ nezamietame :)

2-výberový Studentov t-test

H_0 zamietame ak $\mu_x >> \mu_y$ alebo $\mu_x << \mu_y$ `var.equal=TRUE` → predtest dopadol dobre
p-value $= 7.17\% > 5\%$, teda H_0 nezamietame
→ t-test povedal, rozdiel ($\bar{X} = 856, \bar{Y} = 820.5$) nie je štatisticky významný lebo vzorka 20 dát je málo

2-výberový Welchov t-test

1. a 2. deň $H_0 : \sigma_x^2 = \sigma_y^2 \qquad vs. \qquad H_1 : \sigma_x^2 \neq \sigma_y^2 \qquad \frac{S_x^2}{S_y^2} = 2.9$:(
→ 2-výverový Welchov t-test `var.equal=FALSE`

1. $T = \frac{\bar{X} - \bar{Y}}{\sqrt{\cdots}} \dot{\sim} t_\nu$ (tento test je približný)

2. $\nu = ?$, treba ho odhadnúť (#st. voľnosti)

p-value $= 6\%$, $H_0 : \sigma_x^2 = \sigma_y^2$ nezamietame (dát je málo a majú veľkú varianciu)

Párový t-test

Predpoklad St. t-testu a Welchovho t-testu: dáta X_i a Y_i sú nezávislé

Dáta podrážok: opotrebovanie pred a po nosení, rozdiel je 1 číslo – ∀ dieťa nosilo topánku A aj B, porovnávame opotrebovanie materialA: X materialB: Y

$$H_0 : \mu_x \leq \mu_y \qquad vs. \qquad H_1 : \mu_x > \mu_y \text{ (research hypothesis)}$$

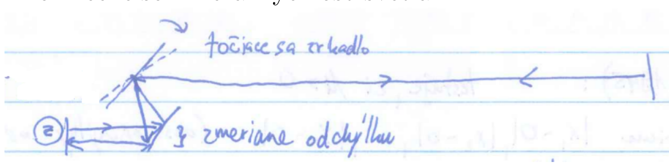
Hypotéza: A sa opotrebováva viac ako B.
Nemôžeme spraviť studentov, welchov t-test, lebo dáta nie sú nezávislé (v stĺpcoch, pravá/ľavá topánka).
→ použijeme párový test → dáta v dvojiciach $\binom{X_1}{Y_1} \ldots \binom{X_n}{Y_n} \sim N_2\left(\binom{\mu_x}{\mu_y}, \begin{pmatrix} \sigma_x^2 & cov \\ cov & \sigma_y^2 \end{pmatrix}\right)$

1. párovanie[2] $Z_i = X_i - Y_i \sim N(\mu_x - \mu_y), (1, -1)\begin{pmatrix} \sigma_x^2 & cov \\ cov & \sigma_y^2 \end{pmatrix}\binom{1}{-1} = \sigma_x^2 + \sigma_y^2 - 2cov)$

2. 1-súborový studentov t-test na dátach $Z_1, \ldots, Z_n$, $(T = \frac{\bar{Z}-0}{S}\sqrt{n} > t_{n-1}(5\%))$

Ak do t-testu dáme závislé dáta, bude sa správať <u>konzervatívne</u>, vysoká p hodnota

Ako Mechelson meral rýchlosť svetla?



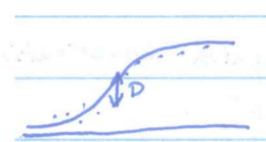
problémy: kým svetlo prešlo tam a naspäť, zrkadlo sa veľmi nestihlo pohnúť
(Poliak) $\to$ rýchlejšie potreboval roztočiť zrkadlo $\to$ použil parný stroj = zdroj nepresností (nestále otáčky)
– prvý Američan, ktorý získal Nobelovu cenu za vedu

Testy normality

Kolmogor-Smirnov test

$X_1, \ldots, X_n \overset{?}{\sim} N(,)$ $\quad H_0$: data $\sim N(\ldots, \ldots)$ $\quad vs. \quad H_1$: data nie su z$N(\ldots, \ldots)$
dáta pochádzajú z rozdelenia, ktoré má <u>distribučnú funkciu</u>[3] $F(\ldots) =?$ (nepoznáme ju)
$\to$ odhad $F(7) = P(X < 7) = \frac{\#\{x_i < 7\}}{n} =: \hat{F}(7) \cdots$ ECDF: Empirical CDF
Idea testu: Ak platí H_0, tak ECDF $\hat{F}(\ldots)$ by sa mala podobať na CDF pre $N(\mu, \sigma^2)$ (pnorm). Za μ zvolíme $\bar{X}$, za σ^2 zvolíme S^2.

H_0 zamietame ak $\hat{F}$ a pnorm$(\bar{x}, S)$ sú veľmi oslišné.
D: maximálna zvislá odchýlka >> 0 (test. štatistika)
p-hodnota: D sa riadi nejakým rozdelením $\to$ Kolmogorov-Smirnov ho zrátali
pre rýchlsoť svetla $p = 45\% >> 5\%$, H_0 nezamietame



Shapirov-Wilkov test (pre malé sady dát)

– snaží sa zistiť, či kvantilový diagram vyzerá ako praimka
Žiaden test normalitu Z_i (podrážky dáta A - dáta B) nezamietal napriek tomu, že histogram naznačuje, že dáta nie sú normálne, ale vznil <u>z 10 dát</u>.

Neparametrické testy

Levi-Strauss – testovali naraz nanášanie farby ľuďmi vs. strojmi an látku
– dáta: % nepodarkov ľudia-stroje $\to$ o koľko % sa ľudia viac mýlili
$\to$ dát je len 22, ale napriek tomu SW test normalitu zamieta $\quad \to$ je tam veľa outlierov, pri normálnom rozdelení by mali byť približne $\frac{1out.}{100dat}$.
$X_1, \ldots, X_n \nsim N()$
$?\mu = E(X)$ – stredná hodnota nepodarkov ľudí voči strojom

$$H_0 : \mu \leq 0 \qquad vs. \qquad H_1 : \mu > 0$$

H_1 = ľudia sú horší ako stroje
–nemôžeme použiť Studentov t-test (nemáme N)

[2] Stačí overiť že rozdiely $Z \sim N$
[3] CDF: Cumulative Distribution Function

Wilcoxon(1945)

testuje, či $\mu > 0$

1. Zoberieme $|X_1 - 0|, |X_2 - 0|, \dots, |X_n - 0|$ (abs. odchýlky od strednej hodnoty)
2. orankujeme odchýlky $\dots n \dots 2 \dots 1 \dots 3 = R_1 \dots R_n$
3. $S_W = \sum_{i, x_i - 0 > 0} R_i$ (súčet rankov, kde bola odchýlka kladná), H_0 zamietame ak $S_W >> 0$.

$X_1, \dots, X_m \nsim N(,)$ a $Y_1, \dots, Y_n \nsim N(,)$

$$H_0 : \mu_x = \mu_y \qquad vs. \qquad H_1 : \mu_x \neq \mu_y$$

sú navzájom nezávislé

1. jedna sada dát: $X_1, \dots, X_m, Y_1, \dots, Y_n$
2. Ranky: $\dots 1 \dots 3 \dots 2 \dots m+n = R_1, \dots R_m, R_{m+1}, \dots, R_{m+n}$
3. $S_W = \sum_{i=1}^{m} R_i$ – sčítame len ranky X-ov

– netreba overovať rovnosť disperzií

Ak do t-testu vložíme nenormálne dáta (majú veľa outlierov), t*test sa správa konzervatívne. Ak test $\sigma_1^2 \stackrel{?}{=} \sigma_2^2$ dopadne tesne k 5% → skúsime Welch → nebude fungovať

B.D.Ú. o 2 roky po Wilcoxonovi: MANN, WHITNET 1974 vymysleli ešte jednoduchší postup, všetky dvojice $X_i \overset{\geq}{\underset{<}{=}} Y_j$, test. štatistika $S_{MN} := \#pripadov\{x_i > y_j\}$

– je ich $m \cdot n$

– Wilcoxonova test. štatistika - MANN, WHITNEY = konštanta

1. Zistiť, čomu sa daná konštanta rovná
2. Dokázať, že naozaj bude rovnaká pre ∀ prípady

štat. môže závisieť od n, m, nebude závisieť od konkrétnych dát

<u>Sestry a rukavice – používajú ich zdravotné sestry pri práci s krvou?</u> tajne ich sledovali

školenie → potom znovu prieskum o 1 mesiac

- 2 mesiace → horšie
- 5 mesiacov → ako keby šklenie ani nebolo

<u>Bernoulliho schéma</u> – opakujeme stále ten istý experiment s pravdep. p (áno/nie) → <u>binomické rozdelenie</u>

$X \sim Bin(n, p)$, X – koľkokrát ich použili?, n - # sledovaní vrámci obdobia, p - svedomitosť sestier - nepoznámne (všetkých/priemerne)

1-stranný test

$X = \sum$kedy nosili, $\hat{p} = \frac{x}{n}$

$n = \#$ pozorovaní

v New Yourku je priem, svedomitosť 20%

$$H_0 : p \leq 0.20 \qquad vs. \qquad H_1 : p > 0.20$$

MZR: H_0 zamietame ak $\hat{p} >> 0.2$, teda $\hat{p} - 0.2 >> 0$, teda $\frac{\hat{p} - 0.2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} > u(5\%)$, vtedy H_0 zamietame.

$X \sim Bin(n, p) \leftarrow$ ďalej: akým rozdelením sa riadi $\hat{p} - 0.2$?

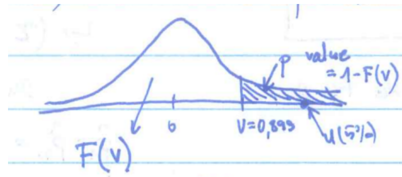
1. LAPLACE-MOIVRE CLT: $\frac{x - np}{\sqrt{np(1-p)}} \dot{\sim} N(0, 1)$
2. MMS: $\frac{\frac{1}{n}}{\frac{1}{n}} \frac{\frac{x}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0, 1)$
3. BDU: treba dokázať $\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0, 1)$

p-value $= 1 - F(v) = 0.18 > 5\% \Rightarrow H_0$ nezamietame

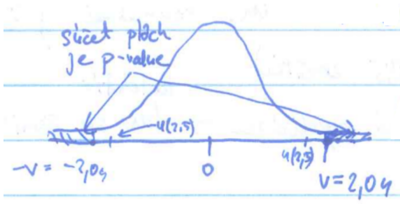
Mali sme tušenie, že priemer našich sestier je väčší ako newyorský.

Ale nie je to štat. významné:

rozdiel 20% a 25% je malý, n = 51 je málo dát



obojstranný test



v USA je priemer svedomitosti 13%

$$H_0 : p = 0.13 \qquad vs. \qquad H_1 : p \neq 0.13$$

MZR: H_0 zamietame, ak $\hat{p} >> 0.13$ alebo $\hat{p} << 0.13$. H_0 zamietame, ak $\frac{\hat{p}-0.13}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} > u(2.5\%)$ alebo $< (2.5\%)$.

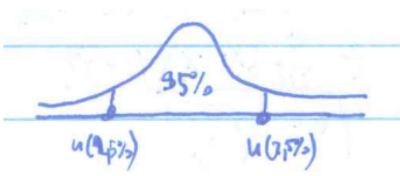
p-value $= 4\% < 5\% \Rightarrow H_0$ zamietame, $H_1 : p \neq 0.13$.

<u>Čo je to p-value?</u>

Čo si ľudia myslia: keď $p < 5\%$, tak H_0 zamietame. AK $p > 5\%$, H_0 nezamietame. H_0 platí nie je náhodá udalosť.

p-value: je pravdepodobnosť, že testovacia štatistika by vyšla ešte väčšia/menšia (extrémnejšia) ako teraz.

Interval spoľahlivosti



$$P(-u(2.5\%) < \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < u(2.5\%)) = 95\%$$

$$P(\hat{p} - u(2.5\%)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + u(2.5\%)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) = 95\%$$

$$L = 0.135 < p < U = 0.374$$

– skutočná svedomitosť $\in (13\%, 37\%)$, $\hat{p} = 0.254$ – stred intervalu

<u>Závisí svedomitosť od skúsenosti?</u>

$X_1 \sim Bin(n_1, p_1 =?)$ najviac 3 roky, $\hat{p}_1 = \frac{7}{10}$

$X_2 \sim Bin(n_2, p_2 =?)$ najviac 3 roky, $\hat{p}_2 = \frac{6}{41}$

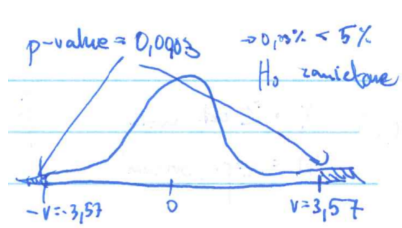
<u>2-sample 2-sided test</u>

$H_0 : p_1 = p_2 \qquad vs. \qquad H_1 : p_1 \neq p_2$

MZR: H_0 zamietame ak $\hat{p}_1 >> \hat{p}_2$ alebo $\hat{p}_1 << \hat{p}_2$

H_0 zamietame ak $\hat{p}_1 - \hat{p}_2 >> 0$ alebo $<< 0$

$V > u(2.5\%)$ alebo $< -u(2.5\%)$



$$V = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

Dá sa odvodiť rovnako ako 1-sample.

Je to štatisticky významné meranie svedmitosti $\rightarrow$ (z 1-str. testu potom zistíme, že menej skúsené sú svedomitejšie)

$\hat{p}_1 - \hat{p}_2 = 0.554$ – bodový odhad

IS $P(-u(2.5\%) < V < u(2.5\%)) = 95\%$

$P(\hat{p}_1 - \hat{p}_2 - \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} u(2.5\%) < p_1 - p_2 < \hat{p}_1 - \hat{p}_2 + sqrt\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2} u(2.5\%))$

IS = (0.249; 0.857) – možno sú len o 25% menej svedomité

$\rightarrow$ prečo vyšiel taký široký? : v prvej skupine(neskúsených sestier) máme len 10 pozorovaní

$X \sim Bin(n,p)$ – je rozumné predpokladať, že každá sestra v nemocnici má rovnaké p?

$\rightarrow$ mohli by sme uvažovať p pre každú sestru zvlášť (logistická regresia – príliš zložité)

$\rightarrow$ je to ale až taká volovina? davová psychóza $\rightarrow$ dav sa zuniformuje

Robert Box: "Every model is wrong, but some are useful."

Korelačná analýza

$$\begin{pmatrix} X \\ Y \end{pmatrix} \begin{matrix} - & INCOME \\ - & FROST \end{matrix}$$

PEARSON corelation coef.

$$\rho := \frac{cov(X,Y)}{\sqrt{D(x)}\sqrt{D(Y)}} = \frac{E((X-E(x))(Y-E(Y)))}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(X-E(X))(Y-E(Y))f(x,y)dxdy}{\sqrt{D(X)}\sqrt{D(Y)}} =?$$

Kde $f(x,y)$ je združená hustota vektora $\binom{X}{Y}$. NIKDY to ale nezrátame.

Dáta $\binom{X_1}{Y_1} \cdots \binom{X_n}{Y_n} \rightarrow$ odhad pre $\rho : \hat{\rho} := \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{S_x^2}\sqrt{S_y^2}} \in < -1, 1 >$.

`cor(Income, Frost) = 0.22` $\rightarrow$ lebo bohatšie regióny sú na S/SV – nie je to kauzalita

`cor(Murder, Life Ex) = -0.87` $\rightarrow$ nie, že by sa vraždili $\Rightarrow$ krátky život/ale je tam zlá situácia (vraždy/krátky živ)

$\rightarrow$ sú to len dohady $\hat{\rho}$ – nerobiť závery len na tom $\rightarrow$ TEST

Testy významnosti korelácie

$H_0 : \rho = 0 \qquad vs. \qquad H_1 : \rho \neq 0$

MZR: H_0 zamietame ak $\hat{\rho} >> 0$ alebo $\hat{\rho} << 0$.

`alternative="two.sided", method="pearson"`

1. ideme testovať, či sú dáta z norm. rozdelenia $\rightarrow$ ks.test

   population & area niu sú z N

2. test.štatistika $\frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}$ porovnáme so t_{n-2} (test závisí aj od počtu dát)

`cor(Income, Frost)` $\rightarrow p - value = 11\% \Rightarrow H_0$ nezamietame $\rightarrow$ vymysleli sme celú teóriu, ale nie je štat. významná

`cor(Income, Hs. Grad)` $\quad H_0 : \rho \leq 0$ vs. $H_1 : \rho > 0$. MZR: H_0 zamietame ak $\hat{\rho} >> 0$.

$\rightarrow p - value = 7e - 7 < 5\% \Rightarrow H_0$ zamietame

$Z =$ population, $Y =$ frost `cor(pop, frost) = -0.33`

$\rightarrow X$ sa neriadia normálnym rozdelením $\rightarrow$ SPEARMAN

SPEARMAN corel. coef

1. dáta $X_1, \dots, X_n \qquad Y_1, \dots, Y_n$
   ranky xov: $\dots n \dots 3 \dots 1 \dots = R_1, R_2, \dots, R_n$
   ranky yov: $\dots 1 \dots n \dots 2 \dots = Q_1, Q_2, \dots, Q_n$

2. $\hat{\rho}_S :=$ pearsonovo $\hat{\rho}$ vyrátaní z rankov

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(R_i - \bar{R})(Q_i - \bar{Q})}{\sqrt{S_R^2}\sqrt{S_Q^2}}$$

   kde $\bar{R}, \bar{Q}, S_R, S_Q$ sú konštanty.

$H_0 : X$ a Y spolu nesúvisia negatívn. vs. $H_1 : X$ a Y spolu súvisia negatívne.
MZR: H_0 zamietame ak $\hat{\rho}_S << 0$
$p - value < 5\% \to H_0$ zamietame $\to$ je štat. významná korelácia Income-Frost
$\hat{\rho} = -0.33 \neq \hat{\rho}_S = -0.46$ ale väčšinou majú rovnaké znamienka
Pearson a Spearman – kolegovia na Cambridge, Pearson pohŕdal Spearmanom[4]
Spearman: Faktorová analýza: výsledok 10boja je určený len 2-3 faktormi: vytrvalosť, rýchlosť, sila

Fischerova Z-premenná

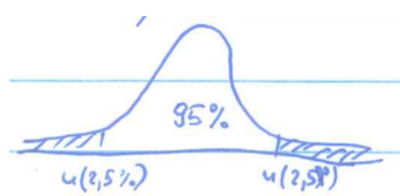
–IS pre ρ ale sú dáta z N(,)

$$\begin{pmatrix} X \\ Y \end{pmatrix} \begin{matrix} - & INCOME \\ - & HS.grad \end{matrix} \qquad H_0 : \rho \leq 0 vs. H_1 : \rho > 0 \dots \text{čomu sa } \rho \text{ rovná?}$$

1. bodový odhad: $\hat{\rho} = 0.61$
2. interval spoľahlivosti pre $\rho \to$ odvodí sa z rozdelenia $\hat{\rho} \dot{\sim} B(\rho, DIVOCINA)$
   - $\dot{\sim}$ platí len pre veľmi veľké n približne :(
   - $DIVOCINA$ závisí od ρ :(
   - použijeme transformáciu: Fischer Z-tranform (hyperbolický arctg.)(ľahký dôkaz Taylorovým rozvojom):

$$atanh(\hat{\rho}) = \frac{1}{2}\ln\frac{1-\hat{\rho}}{1-\hat{\rho}} \dot{\sim} N(atanh(\rho), \frac{1}{n-3}$$

$$\frac{Z - atanh(\rho)}{\sqrt{\frac{1}{n-3}}} \dot{\sim} N(0,1)$$



$$95\% = P(-u(2.5\%) < \frac{Z - atanh/\rho)}{\sqrt{\frac{1}{n-3}}} < u(2.5\%))$$

$$= P(Z - u(2.5\%)\sqrt{\frac{1}{n-3}} < atanh(\rho) < Z + u(2.5\%)\frac{1}{n-3})$$

$$= P(tanh(Z - u(2.5\%)\sqrt{\frac{1}{n-3}}) < \rho < tanh(Z + u(2.5\%)\frac{1}{n-3}))$$

$\to$ tanh je rastúca funckia $\to$ môžeme použiť
$IS = (0.41, 0.76)$ nie je symetrický
POZN: je niečo ako Z-prem pre SPEARMANA $\sim N(\dots, \frac{1.06}{n-3})$, nie pre ľubovoľné rozdelenie

[4] Nebol vyučený štatistik, ale psychológ

2-sample test

$$\begin{pmatrix} X \\ Y \end{pmatrix} \begin{matrix} - & INCOME \\ - & HS.grad \end{matrix}$$

$$JUH: \hat{\rho}_1 = 0.83 \qquad\qquad SEVER: \hat{\rho}_2 = 0.20$$
$$\rho_1 = ? \qquad\qquad \rho_2 = ?$$

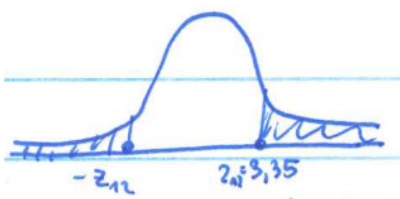
$$H_0: \rho_1 = \rho_2 \qquad vs. \qquad H_1: \rho_1 \neq \rho_2$$

$$Z_1 \dot{\sim} N(atanh(\rho_1), \frac{1}{n_1 - 3}) \qquad Z_2 \dot{\sim} N(atanh(\rho_2, \frac{1}{n_2 - 3})$$

Z_1, Z_2 <u>sú nezavíslé</u> $\Rightarrow$[5] $Z_1 - Z_2 \dot{\sim} N(atanh(\rho_1) - atanh(\rho_2), \frac{1}{n_1-3} + \frac{1}{n_2+3})$

$$\boxed{\frac{(Z_1 - Z_2) - (atanh(\rho_1) - atanh(\rho_2))}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} \dot{\sim} N(0,1)}$$

$\Rightarrow$ TESTOVÁ ŠTATISTIKA $\frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} =: Z_{1,2}$, porovnáme s $u(2.5\%)$.



$p - val \qquad 2(1 - prnom(Z_{1,2})) < 5\% \Rightarrow H_0$ zamietame
TEÓRIA: SEVER: aj nevzdelaní si vedia zarobiť – lepší systém, JUH: nevzdelaní sú kopáči

Lineárna regresia

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2), i = 1 \dots n$
Y_i – ozón,$\beta_0 = ?, \beta_1 = ?$ – parametre regresie, x_i – teplota, ε_i – chyba regres. modelu

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$
$$\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\varepsilon}$$

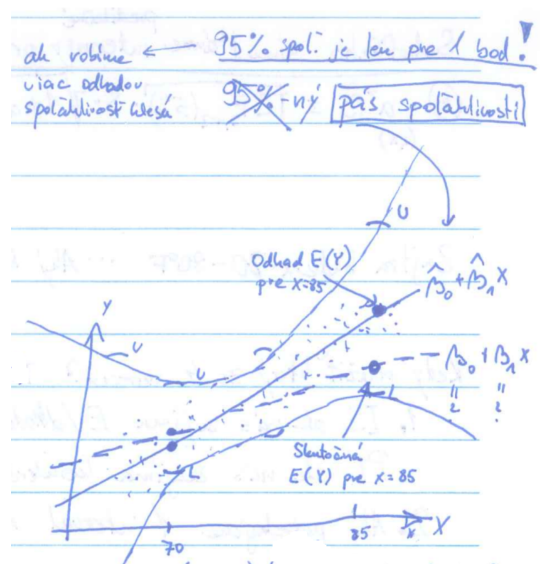
Chceme zistiť $\beta = \binom{\beta_0 = ?}{\beta_1 = ?} \rightarrow$ treba ich odhadnúť.
LS-estimator – LEAST SQUARES $\rightarrow$ súčet zvislých odchýlok = 0, súčet štvorcov.
odchýlky = reziduá $\qquad \hat{\beta} = (X^T X)^{-1} X^T Y$

$$\begin{matrix} \hat{\beta}_0 = \\ \hat{\beta}_1 = \end{matrix} \begin{pmatrix} -2.22 \\ 0.07 \end{pmatrix} = \hat{\beta}$$

KONTRAST: $a_0\beta_0 + a_1\beta_1 = (a_0, a_1)\binom{\beta_0}{\beta_1} = a^T\beta = ?$
ODHAD PRE KONTRAST: $a^T\hat{\beta} \qquad$ PLATÍ $\frac{a^T\hat{\beta} - a^T\beta}{S\sqrt{a^T(X^TX)^{-1}a}} \sim t_{n-2}$
t_{n-2} – n-**2** lebo v modeli sú 2 parametre β
S – odhad pre $\sigma \rightarrow$ smerodajná odchýlka chýb. $S^2 = \frac{SS_e}{n-2}, \qquad SS_e = \sum_{i=1}^n (reziduum)^2 = \sum_{i=1}^n \varepsilon_i^2$, SUM of SQUARES (zase n-2, lebo 2 bety)
$\rightarrow$ z toho dostaneme MMS IS pre β

[5] z nezávislosti aj z predošlého riadku

IS pre $a^T\beta$: (1) $a^T\hat\beta \pm t_{n-2}(2.5\%)\cdot S\sqrt{a^T(X^TX)^{-1}a}$



$$reziduum = \begin{pmatrix} Y_1 - (\hat\beta_0 + \hat\beta_1 x_1) \\ \vdots \\ Y_n - (\hat\beta_0 + \hat\beta_1 x_n) \end{pmatrix} = \vec{Y} - \vec{X}\hat\beta$$

$\beta_1 = ?$ – nepoznáme

1. bodový odhad $\hat\beta_1 = 0.07$ (lin. regres)

2. $(0,1)\binom{\beta_0}{\beta_1} = \beta_1$ – do IS pre kontrast vložíme vektor (0,1)

$x = 85°F(29.4°C) \Rightarrow E(Y) = \ Y$
$E(Y) = E(\beta_0 + \beta_1\cdot 85 + \varepsilon) = \beta_0 + \beta_1 \cdot 85 + E(\varepsilon) = ?$

1. Bodový odhad dosadíme za β_0, β_1 odhady $\hat\beta_0, \hat\beta_1 \Rightarrow \hat\beta_0 + \hat\beta_1\cdot 85 = (1,85)\binom{\hat\beta_0}{\hat\beta_1} = 3.75$, kontrast = (1,85)

2. Intervalový odhad pre $E(Y)$: $(3.61, 3.89) = (L, U)$

   E - str. hodnota – zo zákona veľkých čísel – ak veľakrát zopakujeme pokus, $\bar Y \to E(Y)$

Scheffeho simultanne inetrvaly spoľahlivosti pre $a^T\beta$

(2) $a^T\hat\beta \pm \sqrt{2F_{2,n-2}(5\%)}S\sqrt{a^T(X^TX)^{-1}a}$ – 95% určujú pás spoľahlivosti
Všetky dvojky v $2F_{2,n-2}$ sú # parametrov. F je Fischer-Schnedeker.
rozširovanie na okrahoch: v strede je veľa dát(istejší), na krajoch je menej (menej istý odhad)

Predikčný interval

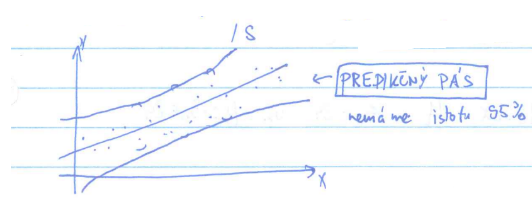
Zajtra bude $85°F$ ... aký bude ozón? $\quad Y = ? = \beta_0 + \beta_1\cdot 85 + \varepsilon$

1. Bodový odhad $\hat\beta_0 + \hat\beta_1 \cdot 85 + 0$

2. Intervalový odhad = PREDIKČNÝ INTERVAL (3) $a^T\hat\beta \pm t_{n-2}(2.5\%)\sqrt{1 + a^T(X^TX)^{-1}a}$

POZN1: PI pre Y[6]: (2.58, 4.92) IS pre $E(Y)$[7]: (3.61, 3.89)
→ Prečo je Pi širší? – suchý dôvod: lebo je vo vzorci 1+, – berie do úvahu ε
POZN2: Čo je predikčný interval?

"chytanie medveďa":
P(teplota zajtra padne do PI) = 95%

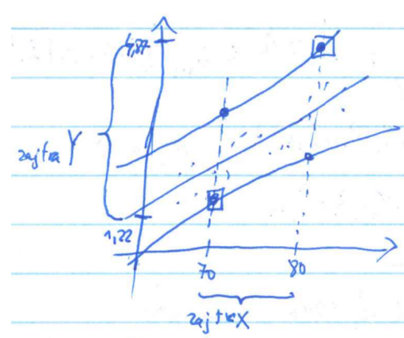


"chytanie motýľa":
P(odhad IS trafí skutočnú hodnotu $E(Y)$)

Scheffeho simultánne predikčné intervaly pre Y

(4) $a^T\hat\beta \pm \sqrt{2F_{2,n-2}(5\%)S\sqrt{1 + a^T(X^TX)^{-1}a}}$

Zajtra bude $x = 70 - 80°F$ ... Aký bude ozón?



[6] $Y = \beta_0 + \beta_1\cdot 85 + \varepsilon \leftarrow$ 3 zdroje neistoty
[7] $E(Y) = \beta_0 + \beta_1 \cdot 85 \leftarrow$ 2 zdroje neistoty

Kedy použiť ktorý zo 4 vzorcov? treba to mať natrénované.

1. IS ak nás zaujíma E/dlhodobý priemer/parameter
   PI ak nás zaujíma konkrétna hodnota
2. Ak potrebujeme 1 inetrval $\rightarrow 1/3$, AK potrebujeme viac intervalov naraz $\rightarrow$ Scheffe

Polynomická regresia

$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$
Kde: $\sigma^2 = ?$, Y = ventilation – ako pumpujú pľúca, x_i – oxygen – koľko kyslíka z neho vedia vytiahnuť.

$$\hat{\beta}_0 = 24.27 \qquad \hat{\beta}_1 = -0.01344 \qquad \hat{\beta}_2 = 0.000008902$$

Vo vzorcoch 1-4 teraz použijeme # parametrov 3. $a = (1, x, x^2)^T$
TEST OF SIGNIFICANCE OF β_2: Chceme zjednodušiť model $\rightarrow$ zanedbať parameter $\hat{\beta}_2$

$$H_0 : \beta_2 = 0 \qquad vs. \qquad H_1 : \beta_2 \neq 0$$

$$\beta_2 = (0,0,1)\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\end{pmatrix}, \qquad a^T = (0,0,1)$$

$$\frac{a^T\hat{\beta} - a^T\beta}{S\sqrt{a^T(X^TX)^{-1}a}} \sim t_{n-3}$$

$$T = \frac{a^T\hat{\beta} - 0}{S\sqrt{a^T(X^TX)^{-1}a}}$$

H_0 zamietame ak $\hat{\beta}_2 >> 0$ alebo ak $\hat{\beta}_2 << 0$.

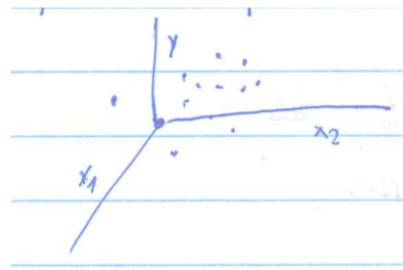
1. $\hat{\beta}_2$ je malé, lebo $x - y$ sú veľké (1000) $\rightarrow x^2$ je $10^6 \rightarrow$ ale β_2 je dôležité
2. β_2 je potrebné, lebo obrázok

Summary

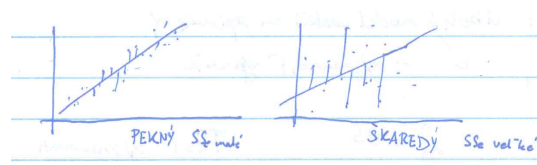
- call – akým príkazom vznikol objekt
- residuals – zvislé odchýlky
- coefficients:
  - stderror – menovateľ T
  - t value – test. štatistika T
  - p-value pre T: $2 \cdot 10^{-16} < 5\% \Rightarrow H_0$ brutálne zamietame
- residual standard error = S, `$sigma`
- Test. štatistiku, že $H_0 : \beta_0 = 24$ $vs.$ $H_1 : \beta_0 \neq 24$ už nedostaneme zo summary $\Rightarrow$ ručne

Porcelán – výsk. ústav zváračský cestou do Rače
– odolnosť porc. voči vysokým teplotám – teplotné šoky teplota1 – 1. šok, potom schladenie, potom 2. šok, potom zmerali pnutie doštičky
–ako teplotné šoky vplývajú na pnutie
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, Y – napätie, x_1 – teplota 1. šoku, x_2 – teplota 2.šoku
$\hat{\beta}_0 = 2.62 \qquad \hat{\beta}_1 = 0.05776 \qquad \hat{\beta}_2 = 0.05667$ je model dobrý?

12 dát, 12 guličiek + prekladáme cez ne rovinu
3D obrázok by sme vedeli nakresliť, 4D si už nechceme ani predstavovať
→ chceme zistiť, či je model dobrý "pomocou číselka"



"krásu" odmeriame pomocou SS_e
SS_e = **miera nekvality modelu**



Aby sme SS_e vedeli porovnať, vytvoríme ÚBOHÝ MODEL(null model) $Y = \beta_0 + \varepsilon$ + ztrátame (total) SS_t
ÚBOHÝ MODEL JE URČITE NEKVALITNEJŠÍ ⇒ $SS_t \geq SS_r$

a) X_1, X_2 sú dôležité na určenie y ⇒ úbohý model bude oveľa horší než náš $SS_t >> SS_e$ $\quad \frac{SS_e}{SS_t} \doteq 0$

$$R^2 = \boxed{1 - \frac{SS_e}{SS_t}} \doteq 1$$

b) X_1, X_2 nie sú dôležité na určenie y ⇒ úbohý model nebude oveľa horší než náš $SS_t \doteq SS_e$ $\quad \frac{SS_e}{SS_t} \doteq 1$

$$R^2 = \boxed{1 - \frac{SS_e}{SS_t}} \doteq 0$$

učenie – determinovanie – R^2 = **KOEFICIENT DETERMINÁCIE** (ako veľmi sú x-y potrebné na určenie Y

summary: multiple R-squared $r.squared

Ale: mnohé znaky nekvality sa pomocou R^2 zistiť nedajú (ale R^2 je najslávnejší, lebo Excel vedel spočítať $\hat{\beta}$ a R^2)

Ktorý šok má vyšší vplyv na napätie . . . ak veríme modelu, je to určené β_1 a β_2

$$H_0 : \beta_1 \leq \beta_2 \qquad vs. \qquad H_1 : \beta_1 > \beta_2$$

MZR: H_0 zamietame ak $\hat{\beta}_1 >> \hat{\beta}_2$, $\hat{\beta}_1 - \hat{\beta}_2 >> 0$, $(0, 1, -1)\hat{\beta} >> 0$. Použijeme test pre kontrasty:

$$T = \frac{a^T \hat{\beta} - 0}{S\sqrt{a^T (X^T X)^{-1} a}}$$

H_0 zamietame ak $T >> 0$.
H_0 sme nezamietli → test pvoedal, že nevieme povedať že β_1 je dôležitejšia ⇒ málo dát, pomerne malý rozdiel

Test významnosti regresie

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \to SS_e$$

$$H_0 : \beta_1 = 0 \wedge \beta_2 = 0 \qquad vs. \qquad H_1 : \beta_1 \neq 0 \vee \beta_2 \neq 0$$

H_1 hovorí, že úbohý nestačí, H_0 hovorí, že úbohý model stačí na popísanie Y
Úbohý mdoel $Y = \beta_0 + \varepsilon \to SS_t$.

H_0 zamietame ak: $SS_t >> SS_e$, teda ak:

$$F = \frac{\frac{SS_t - SS_e}{2}}{\frac{SS_e}{n-3}} >> 0 \qquad F \sim F_{2,n-3}$$

Kde, 2 znamená počet zabitých biet, a š znamená počet biet v pôvodnom modeli. Menovateľ veľkého zlomku je $S^2 = \frac{SS_e}{n-3}$.

$p - value \doteq 5\% \Rightarrow H_0$ zamietame: úbohý mdoel nestačí na popísanie Y

summary: F-statistic = test. štatistika + p-value – významnosť regresie

$Y = \alpha_0 + \alpha_1 t_1 + \alpha_2 t_2 + \varepsilon$, kde Y je napätie, t_1 je trvanie 1. šoku, t_2 je trvanie 2. šoku

$$\hat{\alpha}_0 = 93 \qquad \hat{\alpha}_1 = 0.01992 \qquad \hat{\alpha}_2 = 0.02227$$

Test významnosti: $H_0 : \alpha_1 = 0 \wedge \alpha_2 = 0 \ vs. \ H_1 : \alpha_1 \neq 0 \vee \alpha_2 \neq 0$. $T = 0.92$, $p - value = 41\% \Rightarrow H_0$ nezamietame.
$R^2 = 0.17 << 1$

Testovanie hypotézy o submodeli

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \Rightarrow SS_{e,model}$, kde Y je ozone, x_1 je radiation, x_2 je temp, x_3 je wind.

$$\hat{\beta}_0 = -0.297 \qquad \hat{\beta}_1 = 0.002206 \qquad \hat{\beta}_2 = 0.05 \qquad \hat{\beta}_3 = -0.076$$

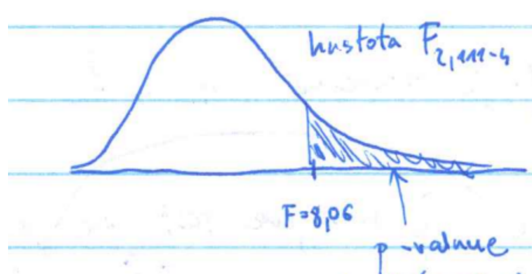
$$H_0 : \beta_0 = 0 \wedge \beta_1 = 0 \qquad vs. \qquad H_1 : \beta_0 \neq 0 \vee \beta_1 \neq 0$$

H_0 – submodel stačí, H_1 – submodel nestačína popísanie Y
Submodel: zohľadňuje $H_0 : Y = \beta_2 x_2 + \beta_3 x_3 + \varepsilon \Rightarrow SS_{e,submodel}$

MZR: H_0 zamietame ak $SS_{e,submodel} >> SS_{e,model}$
$\ldots SS_{e,submodel} - SS_{e,model} >> 0$:

$$F = \frac{\frac{SS_{e,submodel} - SS_{e,model}}{2}}{\frac{SS_{e,model}}{n-4}} \sim F_{2,n-4}$$

$F = 8.06$, $p - value = 1 - pF(2, n-4) = 0.05 < 5\% \Rightarrow H_0$ zamietame – nemôžeme naraz vyhodiť β_0 aj β_1



*alternatívny postup:

1. $H_0' : \beta_0 = 0 \ vs. \ H_1' : \beta_0 \neq 0$ (test hypotézy o kontraste) $\rightarrow$ 5% error typu I.
2. $H_0'' : \beta_1 = 0 \ vs. \ H_1'' : \beta_1 \neq 0$ (test hypotézy o kontraste) $\rightarrow$ 5% error typu I.

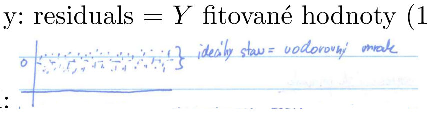
pravidlo: H_0 zamietame ak zamietame H_0' alebo H_0'' $\rightarrow$ chyba I. druhu s $p > 5\%$.
!!! Nedeliť test na viacero, ak sa dá otestovať všetko jedným testom na hladine 5%, použiť ten.

Regresná diagnostika (stručný úvod)

– ako sprojektovať veľarozmerný priestor do $\mathcal{R}^2$, aby sme vedeli zistiť, ako dobre regresia funguje.

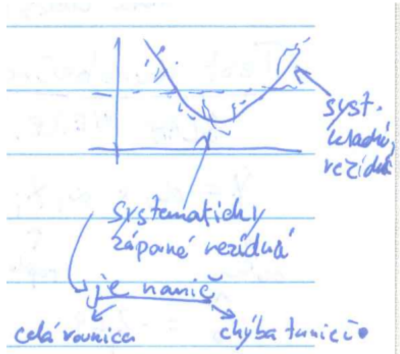
which=1: Residuals vs. Fitted values

– os x: fitted values = odhadované hodnoty $\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i}$ (111 čísel)
– os y: residuals = Y fitované hodnoty (111 čísel)

ideál: 

- označené dni <u>20,23,77</u> majú najvyššie reziduá → sú podozrivé

- ak sú oblasti celé nad/pod → systematické chyby



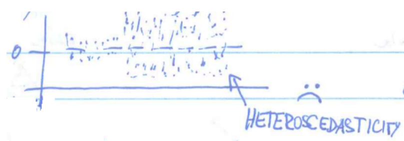
- rozšírené niektoré oblasti

  – veľký rozpor s $\varepsilon_i \sim N(0, \sigma^2)$

  – $\forall \varepsilon$ by mali byť z rovnakého rozdelenia

  $D(\varepsilon_i)$ je konštantná – HOMOSCEDASTICITY

  HETEROSCEDASTICITY– teória je celá založená na konšt. → veľký problém



which=3: Scale-location

– absolútne hodnoty reziduí, odmocníme, znormalizujeme, tiež by mal vyzerať ako vodorovný mrak

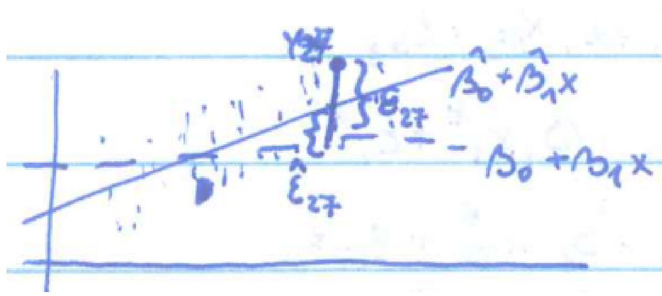
which=2: Normal Q-Q

$\varepsilon_1, \ldots, \varepsilon_n \overset{?}{\sim} N(,)$

– my ich nepoznáme, nie sú to **rezisuá** (ale náh. premenné – odchýlky od skut. regresnej priamky)

–odhadneme ε pomocou $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$ (residuals)

–otestujeme $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n \overset{?}{\sim} N()$ – obrázok / `ks.test`



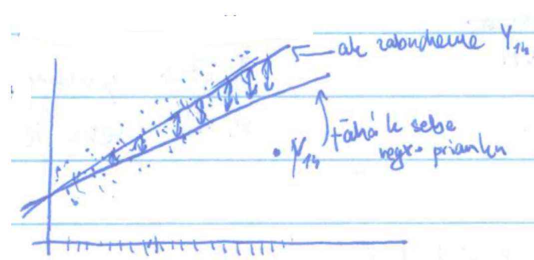
<u>Čo ak zamietame normalitu?</u> ak ε nemajú norm. rozdelenie

1. $\hat{\beta} = LS - ESTIMATOR$ od β ✓– funguje bez ohľadu na normálne rozdenenie ε, len iid
2. IS, pásy, PI, predikčné pásy, testy – ×– toto nefunguje, treba použiť náhradné metódy – nezaručujú 95%

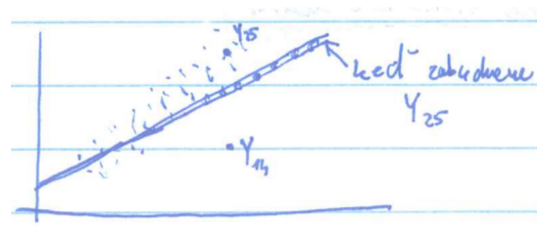
which=4: Cook's distance

– ako veľmi konkrétne body ovplyvňujú regresiu

a) COOK'S DISTANCE$_{14} = \sum_{i=1}^{n} (\updownarrow)^2$
ak bol Y_{14} outlier → je to vysoké číslo



b) COOK'S DISTANCE$_{25} = \sum_{i=1}^{n} (\updownarrow)^2$
ak nebol Y_{25} outlier → je to malé číslo

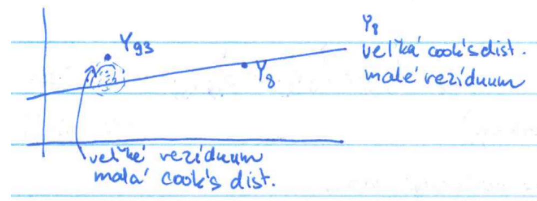


Pre body, ktoré majú vysokú COOK'S DISTANCE treba skúsiť dáta vyhodiť, ak sa veľmi zmenia $\hat{\beta}$ (zmenia znamienka) $\rightarrow$ zmena interpretácie

Podľa Cooka sú podozrivé dáta 17,30,77.

Veľké rezíduum $\not\Rightarrow$ Veľké Cook's distance
Veľké Cook's distance $\not\Rightarrow$ Veľké rezíduum

LEVERAGE POINT (vplyvný bod)
$\rightarrow$ treba hľadať pomocou Cook's dostance



Test rovnobežnosti regresných priamok

SLABÝ VIETOR	SILNÝ VIETOR
$Y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i \qquad (i = 1, \dots, 58)$	$Y_i^* = \beta_0 + \beta_1 x_i^* + \varepsilon_i^* \qquad (i = 1, \dots, 53)$
Y_i – ozón, x_i – teplota	Y_i^* – ozón, x_i^* – teplota
$\hat{\alpha}_0 = -2.69$	$\hat{\beta}_0 = -0.35$
$\hat{\alpha}_1 = 0.078$	$\hat{\beta}_1 = 0.04$

Zdá sa nám, že silný vietor znižuje ozón za rovnakej teploty

$$H_0 : \alpha_1 \leq \beta_1 \qquad vs. \qquad H_1 : \alpha_1 > \beta_1$$

Zložený model

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_{58} \\ -- \\ Y_1^* \\ \vdots \\ Y_{53}^* \end{pmatrix} = \begin{pmatrix} 1 & 0 & x_1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_{58} & 0 \\ -- & -- & -- & -- \\ 0 & 1 & 0 & x_1^* \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & x_{53}^* \end{pmatrix} \cdot \begin{pmatrix} \alpha_0 \\ \beta_0 \\ \alpha_1 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{58} \\ -- \\ \varepsilon_1^* \\ \vdots \\ \varepsilon_{53}^* \end{pmatrix}$$

$$Y = X \cdot \gamma +$$

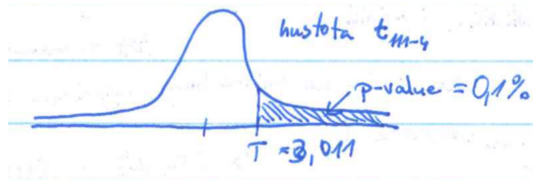
$$\begin{aligned} Y_1 &= \alpha_0 + \alpha_1 x_1 + \varepsilon_1 \\ &\vdots \\ Y_{58} &= \alpha_0 + \alpha_1 x_{58} + \varepsilon_{58} \\ Y_1^* &= \beta_0 + \beta_1 x_1^* + \varepsilon_1^* \\ &\vdots \\ Y_{53}^* &= \beta_0 + \beta_1 x_{53}^* + \varepsilon_{53}^* \end{aligned}$$

Vyjde $\hat{\gamma}^T = (\hat{\alpha}_0, \hat{\beta}_0, \hat{\alpha}_1, \hat{\beta}_1)$.

Máme 1 model, môžeme robiť test:

$$H_0 : \alpha_1 \leq \beta_1 \Leftrightarrow \alpha_1 - \beta_1 \leq 0 \qquad \Rightarrow \text{test o kontraste } a = (0, 0, 1, -1)$$

$$T = \frac{a^T\hat{\gamma} - a^T\gamma}{S\sqrt{a^T(X^TX)^{-1}a}}$$



kde, $a^T\hat{\gamma} = \hat{\alpha}_1 - \hat{\beta}_1$ a $a^T\gamma = 0$.
H_0 zamietame ak $\alpha_1 - \beta_1 >> 0 \Rightarrow T >> 0$.
$p - val = 0.1\% < 5\% \Rightarrow H_0$ zamietame.

Je štatisticky významné, že pri silnom vetre je nižší ozón.